The balanced minimum evolution problem under uncertain data

Catanzaro, Daniele; Labbe, Martine; Pesenti, Raffaele

Published in:
Discrete Applied Mathematics

DOI:
10.1016/j.dam.2013.03.012

Publication date:
2013

Discrete Applied Mathematics 161 (2013) 1789–1804

Contents lists available at SciVerse ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

The balanced minimum evolution problem under uncertain data

Daniele Catanzaro [a,b,*], Martine Labbé [a,1], Raffaele Pesenti [c,2]

[a] Graphes et Optimisation Mathématique, Computer Science Department, Université Libre de Bruxelles (U.L.B.), Boulevard du Triomphe, CP 210/01, B-1050, Brussels, Belgium

[b] Department of Econometrics and Operations Research, University of Groningen, P.O. Box 800, 9700 AV, Groningen, The Netherlands

[c] Department of Management, Universitá Ca' Foscari, San Giobbe, Cannaregio 837, I-30121, Venice, Italy

ARTICLE INFO

Article history:
Received 13 October 2011
Accepted 12 March 2013
Available online 6 April 2013

Keywords:
Network design
Combinatorial optimization
Robust optimization
Bender's decomposition
Computational biology
Balanced minimum evolution
Combinatorial inequalities
Kraft equality

ABSTRACT

We investigate the *Robust Deviation Balanced Minimum Evolution Problem* (RDBMEP), a combinatorial optimization problem that arises in computational biology when the evolutionary distances from taxa are uncertain and varying inside intervals. By exploiting some fundamental properties of the objective function, we present a mixed integer programming model to exactly solve instances of the RDBMEP and discuss the biological impact of uncertainty on the solutions to the problem. Our results give perspective on the mathematics of the RDBMEP and suggest new directions to tackle phylogeny estimation problems affected by uncertainty.

1. Introduction

Let Γ be a set of n objects. A *phylogeny* of Γ is an acyclic graph whose leaves are the objects in Γ and whose internal vertices have degree three [7,8]. For example, by using the convention of denoting the objects in Γ with letters and the internal vertices with numbers, a possible phylogeny of $\Gamma = \{A, B, C, D, E\}$ is shown in Fig. 1.

Consider a symmetric distance matrix D, whose generic entry d_{ij} represents a measure of the dissimilarity between the pair of distinct objects $i, j \in \Gamma$. Then, the *Balanced Minimum Evolution Problem* (BMEP) consists of finding a phylogeny T that minimizes the *length function*

$$L(T, D) = \sum_{i,j \in \Gamma} d_{ij} 2^{-\tau_{ij}(T)} \tag{1}$$

where the *topological distance* $\tau_{ij}(T)$ represents the number of edges belonging to the path from object i to object j in T [10]. For example, with respect to the phylogeny shown in Fig. 1, $\tau_{AB} = 2$, $\tau_{AD} = 4$, and $\tau_{EC} = 3$.

* Corresponding author at: Graphes et Optimisation Mathématique, Computer Science Department, Université Libre de Bruxelles (U.L.B.), Boulevard du Triomphe, CP 210/01, B-1050, Brussels, Belgium. Tel.: +32 2 650 5628, +31 50 363 3794; fax: +32 2 650 5970, +31 50 363 7491.
E-mail addresses: dacatanz@ulb.ac.be (D. Catanzaro), mlabbe@ulb.ac.be (M. Labbé), pesenti@unive.it (R. Pesenti).
1 Tel.: +32 2 650 3836; fax: +32 2 650 5970.
2 Tel.: +39 041 2346927; fax: +39 041 2347444.

0166-218X/$ – see front matter
http://dx.doi.org/10.1016/j.dam.2013.03.012

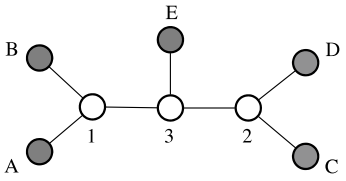

Fig. 1. An example of a phylogeny of five objects $\{A, B, C, D, E\}$ and three internal vertices $\{1, 2, 3\}$.

Solving the BMEP has fundamental practical applications in many research fields, such as medical research, drug discovery, epidemiology, or population dynamics [29]. In these contexts, the objects in Γ are usually called *taxa* and represent molecular data (e.g., DNA, RNA, amino acid or codon sequences) extracted from a given set of species. The values d_{ij} represent *genetic* or *evolutionary distances* between pairs of taxa and a phylogeny of Γ represents the corresponding set of hierarchical evolutionary relationships. These relationships have been of considerable assistance to predict evolution of human influenza A [6], to understand the relationships between the virulence and the genetic evolution of HIV [28,33], to identify emerging viruses as SARS [23], to recreate and investigate ancestral proteins [12], to design neuropeptides causing smooth muscle contraction [3], or to relate geographic patterns to macroevolutionary processes [18].

The $\mathcal{NP}$-hard nature of the BMEP [17] as well as the need of predicting phylogenies from molecular data have justified, in recent years, the development of exact and approximate solution approaches such as those described in [1,10,13,30]. Such approaches rely on point estimations of the evolutionary distances, which are usually computed on the basis of specific substitution models of molecular evolution (see e.g., [11,16]). However, sometimes the lack of biological material, the imprecision of experimental methods, or the combination of unpredictable factors make the evolutionary distances uncertain or difficult to compute. For example, as observed in [15,22], molecular sequences of different genes may contain gaps or missing entries which make them hardly comparable and difficult to align. Moreover, experimental techniques such as DNA or microarray hybridization and comparative serology are limited in terms of pairwise comparisons and often lead to incomplete or uncertain distance matrices [22]. As result, the phylogenies of such taxa are hardly predictable by means of the previous cited solution approaches due to their inability to handle uncertainty in input data.

In this article we show how robust optimization techniques may prove useful to estimate phylogenies from molecular data affected by uncertainty. Specifically, we shall extend Catanzaro et al.'s results [10] by investigating a peculiar version of the BMEP that arises when the evolutionary distances from taxa are uncertain and varying inside intervals $[\underline{d}_{ij}, \overline{d}_{ij}]$, for all $i, j \in \Gamma$. In order to formalize such a version we introduce the following notation. We say that an assignment of possible values to the uncertain distances defines a *scenario* and we denote $\mathbf{D} = \{D \in \mathbb{R}_+^{n\times n} : \underline{d}_{ij} \leq d_{ij} \leq \overline{d}_{ij}, \forall\, i, j \in \Gamma\}$ as the set of all *scenario distance matrices* compatible with a given set Γ. Moreover, we denote $\mathcal{T}$ as the set of all possible $(2n-5)!!$ phylogenies of Γ and, for a fixed scenario $D \in \mathbf{D}$, T_D^* as an *optimal phylogeny* that minimize the length function (1) under the scenario D. Then, a possible way to tackle this specific version of the BMEP consists of minimizing the maximum deviation from the value of an optimal solution over all possible scenarios. This approach is known in literature as *the robust deviation approach* or *the minimax regret approach* [21] and, when used in the context of the uncertain balanced minimum evolution problem, it gives rise to the following combinatorial optimization problem:

Problem 1 (*The Robust Deviation BMEP (RDBMEP)*)**.** Given a set Γ of n taxa and a set of scenario distance matrices $\mathbf{D}$, find the *robust deviation phylogeny*, i.e., the phylogeny $T^\star$ such that

$$T^\star = \arg\min_{T\in\mathcal{T}} \max_{D\in\mathbf{D}} \left[L(T, D) - L(T_D^*, D)\right]. \tag{2}$$

The RDBMEP can be considered as a generalization of the problem investigated in Farach et al. [15] as (i) the set of scenarios considered in the RDBMEP also includes Farach et al.'s ones and (ii) the RDBMEP always has a solution even when Farach et al.'s problem may not. Unfortunately, solving the RDBMEP is at least as difficult as solving the BMEP since the RDBMEP is the robust version of a $\mathcal{NP}$-hard problem [21]. This fact justifies the development of exact and approximate solution approaches to the RDBMEP similar to those proposed by [2,20,24–26,37] for their respective uncertain problems. Hence, in the subsequent sections we shall present a possible mixed integer programming model to exactly solve instances of the RDBMEP and discuss the biological impact of uncertainty on the solutions to the problem.

2. Fundamental properties of the length function

In this section we introduce some fundamental properties of the length function $L(T, D)$ that will prove useful to develop a possible exact approach to solution of the RDBMEP. To this end, we first note that the following proposition holds:

Proposition 1. *For a fixed phylogeny T, the length function $L(T, D)$ is linear in D, it satisfies the condition $L(T, D) - L(T_D^*, D) \geq 0$, and is such that $L(T, D) - L(T_D^*, D)$ is convex in D.*

Proof. The first property trivially holds by the definition of $L(T, D)$. The second property holds by the optimality of the phylogeny T_D^* with respect to D. To prove the third property we need to show that, for any scalar $0 \leq \alpha \leq 1$ and for all $D_1, D_2 \in \mathbf{D}$, it holds that

$$\begin{aligned}&\alpha\left[L(T, D_1) - L(T_{D_1}^*, D_1)\right] + (1-\alpha)\left[L(T, D_2) - L(T_{D_2}^*, D_2)\right]\\ &\quad \geq L(T, \alpha D_1 + (1-\alpha)D_2) - L(T_{\alpha D_1+(1-\alpha)D_2}^*, \alpha D_1 + (1-\alpha)D_2).\end{aligned}$$

By linearity of $L(T, D)$, we have that

$$\alpha L(T, D_1) + (1-\alpha)L(T, D_2) = L(T, \alpha D_1 + (1-\alpha)D_2).$$

Hence, above condition is equivalent to

$$\alpha L(T_{D_1}^*, D_1) + (1-\alpha)L(T_{D_2}^*, D_2) \leq L(T_{\alpha D_1+(1-\alpha)D_2}^*, \alpha D_1 + (1-\alpha)D_2).$$

By exploiting first the optimality of $T_{D_1}^*$ and $T_{D_2}^*$ and subsequently the linearity of $L(T, D)$, it follows that

$$\begin{aligned}\alpha L(T_{D_1}^*, D_1) + (1-\alpha)L(T_{D_2}^*, D_2) &\leq \alpha L(T_{\alpha D_1+(1-\alpha)D_2}^*, D_1) + (1-\alpha)L(T_{\alpha D_1+(1-\alpha)D_2}^*, D_2)\\ &= L(T_{\alpha D_1+(1-\alpha)D_2}^*, \alpha D_1) + L(T_{\alpha D_1+(1-\alpha)D_2}^*, (1-\alpha)D_2)\\ &= L(T_{\alpha D_1+(1-\alpha)D_2}^*, \alpha D_1 + (1-\alpha)D_2),\end{aligned}$$

which concludes the proof. □

With respect to the definition of the RDBMEP, we call the subproblem

$$z(T) = \max_{D\in\mathbf{D}}\left[L(T, D) - L(T_D^*, D)\right] \tag{3}$$

the *Internal Maximization Problem (IMP)* of the RDBMEP. It is worth noting that the IMP is a $\mathcal{NP}$-hard problem as it includes as special case the BMEP. Then, denoted $V\{\mathbf{D}\}$ as the set of the extreme points of $\mathbf{D}$, the following proposition immediately follows from the convexity of function $L(T, D) - L(T_D^*, D)$:

Proposition 2 (*See [34]*)**.** *For any fixed $T \in \mathcal{T}$, there exists an optimal solution to the IMP located on $V\{\mathbf{D}\}$, i.e.,*

$$\max_{D\in\mathbf{D}}\left[L(T, D) - L(T_D^*, D)\right] = \max_{D\in V\{\mathbf{D}\}}\left[L(T, D) - L(T_D^*, D)\right].$$

Proposition 2 states that an optimal solution to the IMP can be found by searching from among the distance matrices D having as generic entry either $\underline{d}_{ij}$ or $\overline{d}_{ij}$. However, it is worth noting that the trivial matrices $\overline{D} = \{\overline{d}_{ij}, \forall(i, j) \in \Gamma\}$ or $\underline{D} = \{\underline{d}_{ij}, \forall(i, j) \in \Gamma\}$ may not be necessarily optimal to (3). In fact, consider the following situation. Let $\Gamma = \{A, B, C, D\}$,

$$\mathbf{D} = \begin{pmatrix} 0 & [1,4] & [2,8] & [2,8] \\ [1,4] & 0 & [2,8] & [2,8] \\ [2,8] & [2,8] & 0 & [1,4] \\ [2,8] & [2,8] & [1,4] & 0 \end{pmatrix}$$

and

$$\hat{D} = \begin{pmatrix} 0 & 4 & 2 & 8 \\ 4 & 0 & 8 & 2 \\ 2 & 8 & 0 & 4 \\ 8 & 2 & 4 & 0 \end{pmatrix}.$$

As $\overline{D}$ is proportional to $\underline{D}$ and both are different from $\hat{D}$, it holds that $T_{\overline{D}}^*$ and $T_{\underline{D}}^*$ are topologically equivalent, but different from $T_{\hat{D}}^*$, i.e., $T_{\overline{D}}^* = T_{\underline{D}}^* \neq T_{\hat{D}}^*$ (see Fig. 2). Now, consider the instance of the IMP obtained when $T = T_{\overline{D}}^*$ and observe that

$$z(T_{\overline{D}}^*) = L(T_{\overline{D}}^*, \hat{D}) - L(T_{\hat{D}}^*, \hat{D}) > L(T_{\overline{D}}^*, \underline{D}) - L(T_{\underline{D}}^*, \underline{D}) = L(T_{\overline{D}}^*, \overline{D}) - L(T_{\overline{D}}^*, \overline{D}) = 0.$$

Hence, at least in this case the solution to the IMP is not the trivial one.

In the next section we shall introduce a mixed integer linear programming model for the IMP whose relaxation will prove useful to develop an exact algorithm for the RDBMEP.

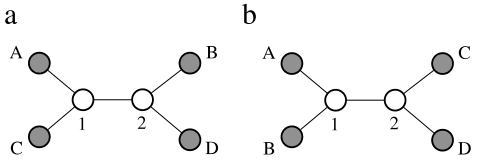

Fig. 2. Graphical representation of the optimal phylogenies $T^*_{\overline{D}}$ (a) and $T^*_{\overline{D}} = T^*_{\underline{D}}$ (b).

3. A mixed integer programming model for the IMP

Given a set Γ, we define $L = \{2, \ldots, n-1\}$ as the set of all the possible values that the topological distances may assume in a phylogeny of Γ; $\tau = \{\tau_{ij} \in L, \forall\, i, j \in \Gamma\}$ as the vector of the unknown topological distances relative to a phylogeny of Γ to be determined; and finally, for a fixed phylogeny T of Γ, $\tau(T) = \{\tau_{ij}(T) \in L, \forall\, i, j \in \Gamma : i < j\}$ as the vector of the topological distances between the taxa of T. For all $i, j \in \Gamma,\ i < j,\ k \in L$, we consider the following decision variable

$$x^k_{ij} = \begin{cases} 1 & \text{if } \tau_{ij} = k \text{ in } T, \\ 0 & \text{otherwise.} \end{cases}$$

We say that variables x^k_{ij} assume *feasible values* if they identify a set of topological distances compatible with a phylogeny in $\mathcal{T}$. Moreover, we denote X as the set of all the possible feasible values for variables x^k_{ij} and we refer the interested reader to [10] for a systematic discussion of the properties that characterize such a set.

In the IMP a phylogeny T and the corresponding vector $\tau(T)$ are given, while the *maximum regret scenario* $D^* = \arg\max_{D \in V\{\mathbf{D}\}} \left[L(T, D) - L(T^*_D, D)\right]$ and the *maximum regret phylogeny* $T^*_{D^*}$ associated to T are unknown. To formulate the IMP as a mixed integer linear programming model we must provide a relationship between the maximum regret scenario distances and the topological distances separating the taxa in Γ on the given phylogeny T. To this end, we first formulate the IMP as double minimization problem as follows:

$$\begin{aligned} \max_{D \in V\{\mathbf{D}\}} \left[L(T, D) - L(T^*_D, D)\right] &= -\min_{D \in V\{\mathbf{D}\}} \left[L(T^*_D, D) - L(T, D)\right] \\ &= -\min_{D \in V\{\mathbf{D}\}} \left[\min_{\tau \in \mathcal{T}} 2\left(\sum_{i<j \in \Gamma} d_{ij} 2^{-\tau_{ij}} - \sum_{i<j \in \Gamma} d_{ij} 2^{-\tau_{ij}(T)}\right)\right] \\ &= -\min_{D \in V\{\mathbf{D}\}} \min_{\tau \in \mathcal{T}} \sum_{i<j \in \Gamma} 2d_{ij}(2^{-\tau_{ij}} - 2^{-\tau_{ij}(T)}). \end{aligned}$$

By replacing τ_{ij} by variables x^k_{ij} we obtain

Formulation 1.

$$\begin{aligned} \min_{D \in V\{\mathbf{D}\}} \min_{\tau \in \mathcal{T}} \sum_{i<j \in \Gamma} 2d_{ij}(2^{-\tau_{ij}} - 2^{-\tau_{ij}(T)}) &= -\min_{D \in V\{\mathbf{D}\}} \min_{x \in X} \sum_{i<j \in \Gamma} 2d_{ij}\left(\sum_{k \in L} 2^{-k} x^k_{ij} - 2^{-\tau_{ij}(T)}\right) \\ &= -\min_{D \in V\{\mathbf{D}\}} \min_{x \in X} \sum_{i<j \in \Gamma} \sum_{k \in L} 2d_{ij}(2^{-k} - 2^{-\tau_{ij}(T)}) x^k_{ij}. \end{aligned} \tag{4}$$

This formulation is valid because, by definition of X (see [10]), for fixed $i, j \in \Gamma, i < j$, only one decision variable x^k_{ij} has value equal to 1, i.e., $x = \{x^k_{ij}\} \in X$ implies $\sum_{k \in L} x^k_{ij} = 1$, for all $i, j \in \Gamma, i < j$. Interestingly, the optimal solution to (4) can be characterized as follow:

Proposition 3. *Formulation* 1 *has an optimal solution* (x^*, D^*) *s.t.*

$$d^*_{ij} = \begin{cases} \overline{d}_{ij} & \text{if } x^{k*}_{ij} = 1 \text{ for some } k > \tau_{ij}(T) \\ \underline{d}_{ij} & \text{if } x^{k*}_{ij} = 1 \text{ for some } k \leq \tau_{ij}(T). \end{cases}$$

Proof. Consider an optimal solution $(x^*, D^*) \in X \times V\{\mathbf{D}\}$ to (4) and assume, by contradiction, that there exists a pair of taxa $i, j \in \Gamma$ such that, for some $k \leq \tau_{ij}(T)$, $d^*_{ij} = \overline{d}_{ij}$. Then, observe that, as the set X is independent of the value of D^*, the solution $(\tilde{x}^*, \tilde{D}^*)$, with $\tilde{x}^* = x^*$ and $\tilde{D}^*$ such that

$$\tilde{d}^*_{rs} = \begin{cases} \underline{d}_{rs} & \text{if } r = i \text{ and } s = j \\ d^*_{rs} & \text{otherwise,} \end{cases}$$

is feasible for (4), i.e., $(\tilde{x}^*, \tilde{D}^*) \in X \times V\{\mathbf{D}\}$. As

$$\sum_{i<j\in\Gamma}\sum_{k\in L} 2d_{ij}^*(2^{-k}-2^{-\tau_{ij}(T)})x_{ij}^{k*} \geq \sum_{i<j\in\Gamma}\sum_{k\in L} 2\tilde{d}_{ij}^*(2^{-k}-2^{-\tau_{ij}(T)})\tilde{x}_{ij}^{k*},$$

then $(\tilde{x}^*, \tilde{D}^*)$ is not worse than (x^*, D^*) which contradicts the initial assumption. It is easy to see that a similar situation also occurs when assuming that, for some $k > \tau_{ij}(T)$, $d_{ij}^* = \underline{d}_{ij}$, hence, the statement follows. □

In lights of Proposition 3 we can further rewrite Formulation 1 in terms of the topological distances separating the taxa in Γ on the given phylogeny, by obtaining the following reformulation:

Formulation 2.

$$\max_{x\in X} \sum_{i<j\in\Gamma}\sum_{k\in L} 2d_{ij}^{k-\tau_{ij}(T)}(2^{-\tau_{ij}(T)}-2^{-k})x_{ij}^k$$

where

$$d_{ij}^{k-\tau_{ij}(T)} = \begin{cases} \underline{d}_{ij} & \text{if } k \leq \tau_{ij}(T) \\ \overline{d}_{ij} & \text{if } k > \tau_{ij}(T). \end{cases}$$

Proposition 4. *Formulation 2 is valid for the IMP.*

Proof. From Eq. (4) and Proposition 3 it follows that

$$\begin{aligned} -\min_{D\in V\{\mathbf{D}\}}\min_{x\in X}\sum_{i<j\in\Gamma}\sum_{k\in L} 2d_{ij}(2^{-k}-2^{-\tau_{ij}(T)})x_{ij}^k &= -\min_{x\in X}\sum_{i<j\in\Gamma}\sum_{k\in L} 2d_{ij}^{k-\tau_{ij}(T)}(2^{-k}-2^{-\tau_{ij}(T)})x_{ij}^k \\ &= \max_{x\in X}\sum_{i<j\in\Gamma}\sum_{k\in L} 2d_{ij}^{k-\tau_{ij}(T)}(2^{-\tau_{ij}(T)}-2^{-k})x_{ij}^k. \quad \square \end{aligned}$$

4. A mixed integer programming model for the RDBMEP

Formulation 2 proves useful to develop a mixed integer programming model for the RDBMEP. Specifically, by using the results presented in the previous section, we can formulate the RDBMEP as follows:

$$\begin{aligned} \min_{T\in\mathcal{T}}\max_{D\in V\{\mathbf{D}\}}(L(T,D)-L(T_D^*,D)) &= \min_{T\in\mathcal{T}}\max_{x\in X}\sum_{i<j\in\Gamma}\sum_{k\in L} 2d_{ij}^{k-\tau_{ij}(T)}(2^{-\tau_{ij}(T)}-2^{-k})x_{ij}^k \\ &= \min_{y\in X}\max_{x\in X}\sum_{i<j\in\Gamma}\sum_{k\in L}\sum_{p\in L} 2d_{ij}^{k-p}(2^{-p}-2^{-k})y_{ij}^p x_{ij}^k \end{aligned} \tag{5}$$

where the parameters d_{ij}^{k-p} are such that

$$d_{ij}^{k-p} = \begin{cases} \overline{d}_{ij} & \text{if } k > p \\ \underline{d}_{ij} & \text{if } k \leq p, \end{cases}$$

and variables $y = \{y_{ij}^p \in \{0,1\}, \forall\, i,j \in \Gamma, \forall p \in L\}$, analogous to variables x in Formulation 2, describe the optimal phylogeny for the RDBMEP in terms of topological distances between taxa in Γ. Following an approach similar to the one already proposed in [27], we can introduce an artificial variable w and rewrite (5) as a standard constrained minimization problem, by obtaining the following Benders' reformulation:

Formulation 3.

$$\begin{aligned} \min_{y,w}\quad & w \\ s.t.\quad & w \geq \sum_{i<j\in\Gamma}\sum_{k,p\in L} 2d_{ij}^{k-p}(2^{-p}-2^{-k})y_{ij}^p\hat{x}_{ij}^k \quad \forall\, \hat{x}\in X \\ & y \in X \\ & w \geq 0. \end{aligned}$$

As vector $\hat{x}$ is constant, Formulation 3 is a linear mixed-integer programming problem on variables y_{ij}^p and w, characterized by an exponential number of constraints, one for each point of X.

Preliminary experiments showed that the solution times of Formulations 2 and 3 are usually very high (over 3 h for instances containing 8–10 taxa), a phenomenon already experienced by [1,10] when tackling instances of the BMEP. To improve this aspect, in the next section we shall describe a possible approach to solution of the RDBMEP inspired by [10], i.e., we shall embody both formulations inside an implicit enumeration algorithm able to simulate the *Stepwise Addition Strategy* (SAS) described in [16].

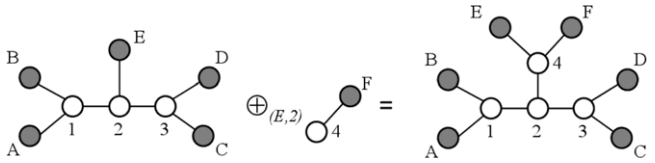

Fig. 3. An example of branching: the phylogeny in the lower part of the figure can be obtained from the one on the top by adding a new edge; in symbols: $Y(\{A, B, C, D, E, F\}) = Y(\{A, B, C, D, E\}) \oplus_{(E,2)} F$.

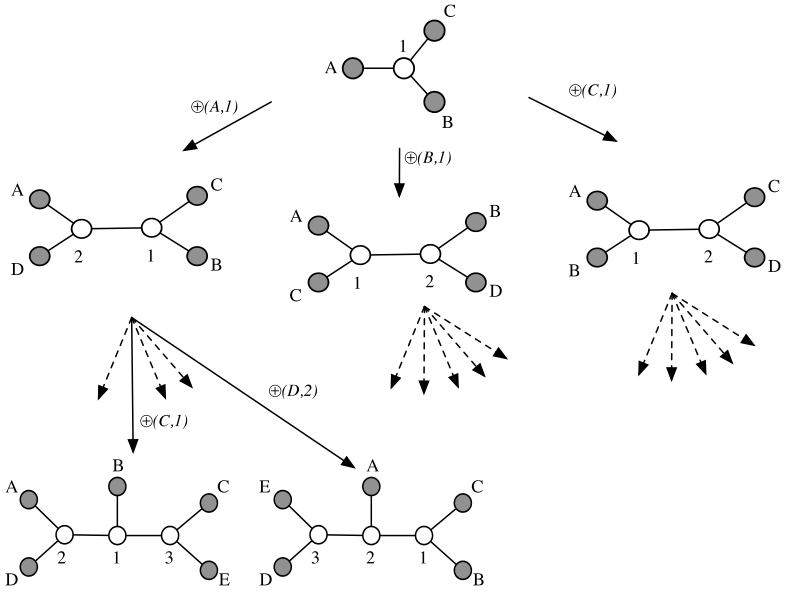

Fig. 4. An example of the first partial phylogenies generated by the implicit enumeration algorithm. Starting from an initial partial phylogeny of three taxa new partial phylogenies are obtained by means of recursive branching operations.

5. Improving the runtime of Formulation 3

Given a phylogeny T of Γ and an taxon $i \in \Gamma$, we denote $\hat{i}$ as the only internal vertex adjacent to i in T. For any subset $S \subseteq \Gamma$, we define a *partial phylogeny* $Y(S)$ as any phylogeny that involves only taxa in S. Denoted $\mathcal{E}(Y(S))$ as the edgeset of $Y(S)$ and considered a taxon $i \in \Gamma \setminus S$ and an edge $(r, s) \in \mathcal{E}(Y(S))$, we define a *branching* as

$$Y(S \cup \{i\}) = Y(S) \oplus_{(r,s)} i = (S \cup \{i\}, (\mathcal{E}(Y(S)) \setminus \{(r, s)\}) \cup \{(r, \hat{i}), (\hat{i}, s), (\hat{i}, i)\}) \tag{6}$$

i.e., as the operation that returns the partial phylogeny $Y(S \cup \{i\})$ obtained by inserting a new edge $(\hat{i}, i)$ on the edge (r, s) of $Y(S)$ (see e.g., Fig. 3). We say that a phylogeny T is *generated* from $Y(S)$ if T is obtained by a recursive branching of $Y(S)$. Then, a possible approach to solution of the RDBMEP consists of: setting S to the subset constituted by the first three taxa in Γ; building the unique partial phylogeny $Y(S)$ of S (see top of Fig. 4); and branching recursively on $Y(S)$ by generating implicitly all possible phylogenies in $\mathcal{T}$ (see Fig. 4) and by computing for each of them the optimal value of the IMP. Hence, the phylogeny minimizing $z(T)$, for all $T \in \mathcal{T}$, will be the optimal solution to the RDBMEP. In this approach, Formulations 2 and 3 play an important role in providing bounds to the respective problems. Specifically, these bounds can be obtained by relaxing in both formulations the integrality conditions for the decision variables and by considering a superset Υ of $C(X)$ defined by the following constraints:

$$x_{ij}^k \geq 0 \quad \forall\, i, j \in \Gamma,\ i < j,\ k \in L \tag{7a}$$

$$\sum_{k \in L} x_{ij}^k = 1 \quad \forall\, i, j \in \Gamma,\ i < j \tag{7b}$$

$$\sum_{k \in L} \left(\sum_{j \in \Gamma : j < i} \frac{x_{ji}^k}{2^k} + \sum_{j \in \Gamma : i < j} \frac{x_{ij}^k}{2^k} \right) = \frac{1}{2} \quad \forall\, i \in \Gamma \tag{7c}$$

$$\sum_{i > j \in \Gamma} \sum_{k \in L} \frac{k x_{ji}^k}{2^k} + \sum_{i < j \in \Gamma} \sum_{k \in L} \frac{k x_{ij}^k}{2^k} = (2n - 3) \tag{7d}$$

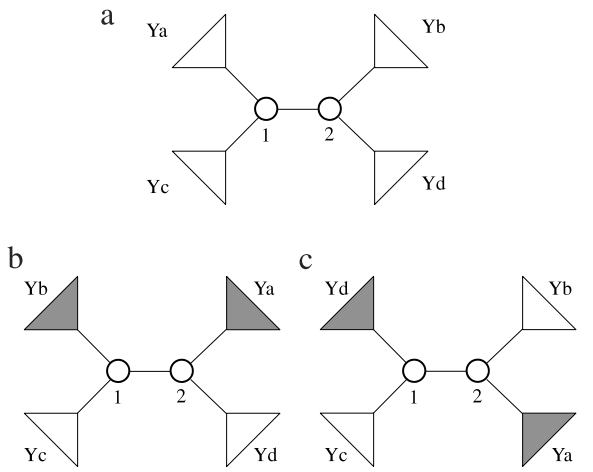

Fig. 5. Examples of Nearest Neighbor Interchange (NNI) on a given phylogeny (shown in figure a) with subtrees Y_A, Y_B, Y_C, Y_D: interchange of subtrees Y_A and Y_B (figure b) and interchange of subtrees Y_A and Y_D (figure c).

where constraints (7a) and (7b) impose general conditions on the topological distances such as the non-negativity and the unicity, respectively. In contrast, constraints (7c) and (7d) impose peculiar conditions that characterize the topological distances of a phylogeny. Specifically, constraint (7c) imposes the so called *Kraft equality* (see [31]), whereas constraint (7d) imposes the *third equality* introduced in [10]. A systematic discussion of the properties and the fundamental equalities that characterize the set X is out of the scope of the present article and can be found in [10].

In order to describe more in detail the solution approach, consider a scenario distance matrix D and a phylogeny T. We define the *regret value* of T associated to T_D^* as the difference $L(T, D) - L(T_D^*, D)$. Moreover, we define the *maximum regret value* of T as the solution value of the internal maximization problem

$$z(T) = \max_{D \in V\{\mathbf{D}\}} \{L(T, D) - L(T_D^*, D)\}.$$

Then, the approach to solution of the RDBMEP can be seen as the result of the interaction of three different algorithms, namely: a local search heuristic, the solver of the IMP, and the solver of the RDBMEP. In the remaining part of this section we shall discuss in details each of them.

The local search heuristic is helpful to provide quick primal bounds to the RDBMEP. In order to describe it, consider two distinct scenarios D^1 and D^2 in $V\{\mathbf{D}\}$. We say that D^1 and D^2 are *adjacent* if there exist two taxa r and $s \in \Gamma$ such that, for all i and $j \in \Gamma$ with $i \neq r$ or $j \neq s$, it holds that $d_{ij}^1 = d_{ij}^2$, while, for $i = r$ and $j = s$ it holds that either $d_{rs}^1 = d_{sr}^1 = \underline{d}_{rs}$ and $d_{rs}^2 = d_{sr}^2 = \overline{d}_{rs}$ or $d_{rs}^1 = d_{sr}^1 = \overline{d}_{rs}$ and $d_{rs}^2 = d_{sr}^2 = \underline{d}_{rs}$. Given a scenario D, we denote $\Delta(D)$ as a set of scenarios constituted by D and its adjacent scenarios.

Similarly to [35] we say that, given two phylogenies $T_1, T_2 \in \mathcal{T}$, T_2 is neighbor of T_1 if T_2 can be obtained by T_1 by means of a single *Nearest Neighbor Interchange (NNI)*, i.e., an operation by which the positions of two subtrees of T_1, whose roots have a topological distance equal to three, are exchanged (see Fig. 5). [35] proved that, starting from a given phylogeny T, the whole set $\mathcal{T}$ can be obtained by applying recursively NNI exchanges on T. Given a phylogeny $T \in \mathcal{T}$, we denote $\mathcal{N}_1(T)$ as the set of phylogenies that can be obtained from T by applying a single NNI exchange. Moreover, we denote $\mathcal{N}_2(T)$ as the set of the phylogenies that are optimal solutions to the problems $T_D^o = \arg\min_{\hat{T} \in \mathcal{T}} L(\hat{T}, D)$, for all $D \in \Delta(D^*)$, being $D^* = \arg\max_{D \in V\{\mathbf{D}\}} \left[L(T, D) - L(T_D^*, D)\right]$. Then, a possible way to approximate the RDBMEP consists of using a local search that iteratively moves from a candidate optimal solution T to a better one by searching both in $\mathcal{N}_1(T)$ and $\mathcal{N}_2(T)$. Specifically, our local search is outlined in Algorithm 1 and alternates two main phases: an *intensification phase* and a *diversification phase*. In the intensification phase (lines 5–8 of Algorithm 1), the local search iteratively searches in the neighborhood $\mathcal{N}_1(T)$ of a given candidate optimal solution T until either a new locally optimal phylogeny is found or no further improvement in $\mathcal{N}_1(T)$ is possible. Subsequently, Algorithm 1 starts the diversification phase (lines 10–13) in which the local search is performed in $\mathcal{N}_2(T)$. When even the diversification phase converges to a locally optimal solution, the local search restarts from the intensification phase and iterates the above steps until no further improvements is possible.

The presence of a diversification phase helps in preventing the premature convergence of Algorithm 1 to poor suboptimal solutions. In fact, as the phylogenies in $\mathcal{N}_2(T)$ are expected to be topologically different from the one obtained at the end of the intensification phase (since they are, by definition, the phylogenies that maximize the regret $L(T, D) - L(T_D^o, D^*)$), Algorithm 1 is forced to search in different (potentially all) subsets of the search space $\mathcal{T}$. The computation overhead required at lines 3, 6, 8, 11, and 10 in Algorithm 1 may be heavy, as such lines involve solving two $\mathcal{NP}$-hard problems (i.e., the IMP and the determination of the elements of $\mathcal{N}_2(T)$). To speed up computation, the local search uses greedy heuristics to tackles both problems. Specifically, as regards the IMP, the local search does not compute the optimal value $z(T)$ but the optimal

Algorithm 1: Local Search Algorithm

```
LocalSearch(Γ, T);
input  : Γ: a set of taxa
         T: a first candidate optimal phylogeny, possibly NULL
output : A phylogeny suboptimal solution to the RDBMEP
// Initialization - if T = NULL Determine a first candidate optimal solution;
if T = NULL then
1     D_0 = random element of V{D};
2     T = arg min_{T∈𝒯} L(T, D);
3  z = max_{D∈V{D}} {L(T, D) − L(T*_D, D)};
4  repeat
       //Intensification phase - Search for a better solution in 𝒩_1(T);
5      forall the T̃ ∈ 𝒩_1(T) do
6          z̃ = max_{D∈V{D}} {L(T̃, D) − L(T*_D, D)};
7          if z̃ < z then
8              z = z̃; T = T̃; D = arg max_{D∈V{D}} {L(T̃, D) − L(T*_D, D)};
9              break;
       //Diversification phase - Search for a better solution in 𝒩_2(T);
10     forall the T^o_D ∈ 𝒩_2(T) do
11         z̃ = max_{D∈V{D}} {L(T^o_D, D) − L(T*_D, D)};
12         if z̃ < z then
13             z = z̃; T = T^o_D;
   until the solution (T, D) has improved;
   return T
```

value of the relaxed version of Formulation 2 previously discussed. In addition, at line 8 of Algorithm 1, instead of computing $D^* = \arg\max_{D\in V\{\mathbf{D}\}}\{L(\tilde{T}, D) - L(T^*_D, D)\}$, the value of D^* is determined by rounding to the closest extreme of $\mathbf{D}$ the vector $\{\sum_{k\in L} d_{ij}^{k-\tau_{ij}(T)} x_{ij}^{k*},\ \forall\, i, j \in \Gamma\}$, where $\{x_{ij}^{k*},\ \forall\, i, j \in \Gamma,\ \forall k \in L\}$ is the optimal solution of the relaxation of Formulation 2. Finally, as regards the determination of the set $\mathcal{N}_2(T)$, the local search uses the NNI heuristic (described in [14]) to quickly find a suboptimal solution to the problems $T^o_D = \arg\min_{\hat{T}\in\mathcal{T}} L(\hat{T}, D)$, for all $D \in \Delta(D^*)$.

The subroutine SolveIMP constitutes the main algorithm that solves the IMP. Given a set Γ of taxa, a phylogeny $T \in \mathcal{T}$ and, optionally, an estimation z on the maximum regret value of T, the subroutine SolveIMP, outlined in Algorithm 2, returns the maximum regret value $z^* = z(T)$ of T if $z^* > z$, and z otherwise. The subroutine implicitly enumerates all the phylogenies in $\mathcal{T}$, in search of a maximum regret phylogeny of T and while doing that it confronts their associated regret values with z. SolveIMP implements the *Stepwise Addition Strategy* (SAS) described in [16] to enumerate the phylogenies in $\mathcal{T}$. Specifically, at lines 1–2 in Algorithm 2, the subroutine first creates a partial phylogeny of three taxa (see, e.g., the partial phylogeny made of three taxa at the top of Fig. 4). Then, in line 3, SolveIMP calls the subroutine SearchMax, described below, that recursively performs branching operations (6) on the partial phylogeny to potentially generate all the possible phylogenies which are solutions to the problem.

The subroutine SearchMax is the core of the SAS algorithm. SearchMax recursively branches on a given partial phylogeny $Y(S)$ in order to find the maximum regret phylogeny associated to T better than the current one $\hat{T}$. The subroutine stops when either one of the following situations occurs:

i. A phylogeny having an associated regret value greater than the one associated to the current maximum regret phylogeny $\hat{T}$ is found. In this case SearchMax returns the new phylogeny and its associated regret value.
ii. It can be proved that no phylogeny generable by $Y(S)$ has an associated regret value greater than $\hat{T}$. In this case SearchMax returns $\hat{T}$ and z.

The second situation is critical, as it implies the existence of a bounding function able to decide whether to branch or stop. In the subroutine SearchMax this task is carried out by function BoundMax at line 1 of Algorithm 3. Specifically, if $S \subset \Gamma$, BoundMax returns an upper bound $\tilde{z}$ on the maximum regret value of T associated to the phylogenies generable by $Y(S)$.

Differently, if $S = \Gamma$, BoundMax returns the exact regret value $\tilde{z}$ of T associated to $Y(S)$. Note that, if $\tilde{z} \leq z$, there is no hope to find a phylogeny generated by $Y(S)$ with associated regret value greater than the current one, hence SearchMax stops the recursion. If $\tilde{z} > z$ and $S = \Gamma$, a phylogeny having regret value greater than the current one has been found, hence SearchMax updates the maximum regret phylogeny and stops the recursion (see lines 3–4 in Algorithm 3). Finally, if $\tilde{z} > z$ but $S \subset \Gamma$, nothing can be said so that SearchMax keeps branching on $Y(S)$ until a phylogeny of Γ is determined (see lines 5–7).

Given a phylogeny $T \in \mathcal{T}$, BoundMax computes an upper bound $\tilde{z}$ on the maximum of the regret values of T associated to the phylogenies generable by $Y(S)$. Specifically, it solves the relaxation of Formulation 2 with an additional set of constraints

Algorithm 2: Exact solver for the IMP

```
SolveIMP(Γ, T, z);
input : Γ: a set of taxa;
        T: a phylogeny for Gamma ;
        z: an esteem of the value of the IMP solution;
output: z*: value of the optimal solution of the IMP (3);
// Initialization of the SAS algorithm;
1 set S = {Head(1, Γ), Head(2, Γ), Head(3, Γ)};
2 let Y(S) be the unique partial phylogeny of S;
// recursive part of the SAS algorithm;
3 (T*, z*) = SearchMax(S, Y(S), NULL, T, z);
return z*;
```

Algorithm 3: SearchMax algorithm

```
SearchMax(S, Y(S), T̂, T, z);
Input : S: a subset of taxa
        Y(S): a partial phylogeny on S
        T̂: the current tentative maximum regret phylogeny for T;
        T: a phylogeny for Γ
        z: an esteem of the value of the IMP solution;
Output: T̂: a tentative maximum regret phylogeny;
        z: the regret value associated to T̂
1 if (z̃ = BoundMax(S, Y(S), T)) > z then
2     if (S == Γ) then
3         T̂ = Y(S);
4         z = z̃;
      else
5         set i = Head(|S| + 1, Γ);
6         for e ∈ E(T(S)) do
7             SearchMax(S, Y(S) ⊕_e i, T̂, T, z);
return (T̂, z);
```

Algorithm 4: Exact Solver for the RDBMEP

```
ExactAlg(Γ);
input : Γ: the set of taxa
output: An phylogeny T* optimal solution to the RDBMEP and its associated value z* = f(T)
// Initialization - Determine a first candidate optimal phylogeny;
1 T = LocalSearch(Γ, NULL);
2 z =SolveIMP(Γ, T, NULL);
// Recursive search of an optimal phylogeny - SAS algorithm;
3 set S = {Head(1, Γ), Head(2, Γ), Head(3, Γ)};
4 let Y(S) be the unique partial phylogeny of S;
5 (T*, z*) = Search(S, Y(S), T, z);
return (T*, z*);
```

Algorithm 5: SearchMin algorithm

```
SearchMin(S, Y(S), T, z);
Input : S: a subset of taxa
        Y(S): a partial phylogeny
        T: current optimal phylogeny
        z: the value of the candidate optimal phylogeny
Output: T: a tentative optimal phylogeny;
        z: the optimal value associated to T
1 if BoundMin(S, Y(S)) < z then
2     if (S == Γ) and ((z̃ =SolveIMP(Γ, Y(S), z)) < z) then
3         T = Y(S);
4         z = z̃;
      else
5         set i = Head(|S| + 1, Γ);
6         for e ∈ E(T(S)) do
7             Search(S, Y(S) ⊕_e i, T, z);
return (T, z);
```

that excludes any phylogeny that cannot be generated by $Y(S)$. By using the same notation of [10], the additional set of constraints can be stated as follows. Given two distinct taxa q and $t \in S$, let σ_{qt} be the topological distance between taxa q

and t in $Y(S)$. Then we have the following conditions on the topological distances between the pair of taxa on the phylogenies that can be generated by $Y(S)$:

i. For all i and $j \in S$, $i < j$,

$$\sigma_{ij} \leq \tau_{ij} \leq \sigma_{ij} + |\Gamma \setminus S|. \tag{8}$$

These inequalities hold as, on each of the remaining $|\Gamma \setminus S|$ branchings needed to obtain a complete phylogeny for Γ, the distance between i and j cannot decrease, and it increases by one only if the branched edge is on the paths between i and j. In terms of the x_{ij}^k variables, the above constraints become $x_{ij}^k = 0$ for all $k < \sigma_{ij}$ and all $k > \sigma_{ij} + |\Gamma \setminus S|$.

ii. For all $i \in S$ and $j \in \Gamma \setminus S$, $i < j$,

$$2 \leq \tau_{ij} \leq \max\left(\max_{q \in S: i > q}\{\sigma_{qi}\}, \max_{q \in S: i < q}\{\sigma_{iq}\}\right) + |\Gamma \setminus S|. \tag{9}$$

Specifically, $\tau_{ij} = 2$ is achieved when edge $(\hat{j}, j)$ is inserted on the edge $(\hat{i}, i)$ and the two edges are not branched any more. Differently, $\tau_{ij} = \max\left(\max_{q \in S: i > q}\{\sigma_{qi}\}, \max_{q \in S: i < q}\{\sigma_{iq}\}\right) + |\Gamma \setminus S|$ is achieved when $(\hat{j}, j)$ is inserted on the edge $(\hat{q}^*, q^*)$, being $q^* = \arg\max\left(\max_{q \in S: i > q}\{\sigma_{qi}\}, \max_{q \in S: i < q}\{\sigma_{iq}\}\right)$, and the subsequent branchings are always performed on an edge belonging to the path between i and j. In terms of the x_{ij}^k variables, the above constraints become $x_{ij}^k = 0$ for all $k > \max\left(\max_{q \in S: i > q}\{\sigma_{qi}\}, \max_{q \in S: i < q}\{\sigma_{iq}\}\right) + |\Gamma \setminus S|$.

iii. For all i and $j \in \Gamma \setminus S$, $i < j$,

$$2 \leq \tau_{ij} \leq \max_{t < q \in S}\{\sigma_{tq}\} + |\Gamma \setminus S|. \tag{10}$$

Specifically, $\tau_{ij} = 2$ is achieved when edge $(\hat{j}, j)$ is inserted on the edge $(\hat{i}, i)$ and the two edges are not branched any more. Differently, $\tau_{ij} = \max_{t < q \in S}\{\sigma_{tq}\} + |\Gamma \setminus S|$ is achieved when: $(\hat{i}, i)$ is inserted on the edge $(\hat{t}^*, t^*)$; $(\hat{j}, j)$ is inserted on the edge $(\hat{q}^*, q^*)$, being $(t^*, q^*) = \arg\max_{t < q \in S}\{\sigma_{tq}\}$; and the subsequent branchings are always performed on an edge belonging to the path between i and j. In terms of the x_{ij}^k variables, the above constraints become $x_{ij}^k = 0$ for all $k > \max_{t < q \in S}\{\sigma_{tq}\} + |\Gamma \setminus S|$.

When $S = \Gamma$ the above bounds trivially reduce to $\tau_{ij} = \sigma_{ij}$, i.e., $x_{ij}^k = 1$, for $k = \sigma_{ij}$, and 0 otherwise, for all $i, j \in \Gamma$, $i < j$. Hence, when $S = \Gamma$ the relaxation of Formulation 2 with the above additional constraints returns the exact value of $\max_{D \in V\{\mathbf{D}\}}(L(T, D) - L(Y(S), D))$.

Solving the RDBMEP. The subroutines before described can be combined to design an implicit enumeration algorithm, called ExactAlg and outlined in Algorithm 4, that exactly solves the RDBMEP. Specifically, ExactAlg first calls the heuristic algorithm LocalSearch to obtain a starting phylogeny T. Subsequently, at line 2, ExactAlg computes the maximum regret value $z = z(T)$ of T calling the subroutine SolveIMP. Finally, at lines 3–5, ExactAlg implicitly enumerates all phylogenies in $\mathcal{T}$ by using z as a first upper bound on the value of the optimal phylogeny. Once again, the implicit enumeration exploits the SAS algorithm, in a way similar to the one described before for the SolveIMP. Specifically, ExactAlg first creates a partial phylogeny of three taxa at lines 3–4. Then, in line 5, ExactAlg calls the recursive subroutine SearchMin, outlined in Algorithm 5, that behaves in a way similar to the subroutine SearchMax. Hence, the whole algorithm can be imagined as the result of the interaction of two nested recursive SASs, one used to solve the minimization problem and one for the internal maximization problem.

SearchMin and SearchMax differ only for the bound functions used. Specifically, SearchMin, at line 1, calls the function BoundMin to compute a lower bound on the maximum regret values of the phylogenies generable by $Y(S)$ by solving the relaxation of Formulation 3 with the additional constraints (8)–(10). As Formulation 3 may include an exponential number of constraints

$$w \geq \sum_{i<j \in \Gamma} \sum_{k,p \in L} 2d_{ij}^{k-p}(2^{-p} - 2^{-k}) y_{ij}^p \hat{x}_{ij}^k \quad \forall \hat{x} \in X. \tag{11}$$

BoundMin does not consider all of them at each branching process of the SearchMin algorithm. On the contrary, BoundMin uses Bender's decomposition to implement (11). Specifically, BoundMin updates the relaxation of Formulation 3 by gradually increasing the subset of constraints of type (11) taken into account. Initially, such a subset is constituted by just a random sample of elements $\hat{x} \in X$. As rule of thumb we usually pick 100 samples for each taxon in Γ. Then, each time that SearchMin calls SolveIMP to compute $z(Y(S))$, if it holds that $\tilde{z} < z$, a new element $\tilde{x} \in X$ is included in the subset, being $\tilde{x}$ the binary vector that describes the topological distances among the taxa in Γ on the maximum regret phylogeny individuated by subroutine SolveIMP. This strategy is similar to the one described in [24] for the robust spanning tree problem. As we do not consider all the constraints of type (11) the subroutine BoundMin does not return the maximum regret value of the phylogeny $Y(S)$, when $S = T$. In fact, we try to avoid as much as possible the evaluation of the maximum regret value of $Y(S)$ by using SolveIMP, as this task is computationally hard. We call SolveIMP only if the lower bound computed through BoundMin does not cut off $Y(S)$.

6. Computational results

We tested the performances of our algorithm on the same real aligned DNA datasets presented in [1,9,10], namely: "Primates12/898", a dataset of 12 sequences, 898 characters each from primates mitochondrial DNA; "RbcL55/1314", a dataset of 55 sequences, 1314 characters each of the rbcL gene; "Rana64 /1976", a dataset of mitochondrial DNA containing 64 taxa of 1976 characters each from ranoid frogs; "M17/2550", "M43/2086", "M18/8128", "M82/2062", "M62/3768", five datasets of respectively 17 sequences of 2550 characters each from insects, 43 sequences of 2086 characters each from mammals, 18 sequences of 8128 characters each from cetacea, 82 sequences of 2062 characters each from fungi, and 62 sequences of 3768 characters each from hyracoidae; finally, "SeedPlant25/19784", a dataset of 25 sequences of 19 784 characters each from pinoles.

From each dataset we extracted the first 20 taxa (or all taxa if $n < 20$) and built the associated $n \times n$ distance matrices by using the General Time Reversible (GTR) model of DNA sequence evolution in which all the gaps were treated as 'N'. The estimation method used to obtain GTR distances is described in [11]. The instances used in [1,9,10] are deterministic. To simulate the presence of uncertainty, we considered possible three ranges of variations for the distances, namely: 5%, 10%, and 15%. Specifically, for each input distance matrix $D = \{d_{ij}\}$ we built a new instance of the RDBMEP by replacing the entries of D by the intervals $[\underline{d}_{ij}, \overline{d}_{ij}] = [d_{ij}(1-\mu), d_{ij}(1+\mu)]$, $i, j \in \Gamma$, where μ is a parameter uniformly distributed in $[0, \alpha]$ and $\alpha \in \{0.05, 0.1, 0.15\}$. Subsequently, in order to correlate the dimension of the instances to the solution time necessary to exactly solve them, we extracted from each instance of the RDBMEP the corresponding k-th leading principal submatrices, $k \in [10, \ldots, \max]$, where max is 12 for Primates12, 17 for M17, 18 for M18, and 20 for the remaining datasets. We implemented our algorithm in ANSI C++ by using Xpress Optimizer libraries v18.10.00. The experiments run on a Pentium 4, 3.2 GHz, equipped with 2 GB RAM and operating system Gentoo release 7 (kernel linux 2.6.17). We deactivated the Xpress pre-solving strategy, assumed three hours as maximum runtime per instance, and used the Xpress dual simplex to compute the linear programming relaxations of the formulation presented in the previous sections. The code and the datasets used in our experiments can be downloaded at http://homepages.ulb.ac.be/dacatanz/Site_4/Software.html.

Table 1 summarizes the results obtained in our experiments. Specifically, fixed an uncertainty level, the column "Optimum" refers to the optimal value to a specific instance. The column "Time" refers to the solution time (expressed in seconds) taken to solve exactly a specific instance. The column "Gap" (expressed in percentage) refers to the difference between the optimal value to a specific instance and the value of linear relaxation of Formulation 3 at the root node of the search tree, divided by the optimal value. Finally, the column "Nodes" refers to the number of nodes needed to solve exactly a specific instance. The symbol n.a. denotes those instances for which the computation took more than 3 h. The table shows that already with an uncertainty level of 5% the instances of the RDBMEP become much harder to solve with respect to the analogous deterministic instances of the BMEP (see [10]). In fact, a part from Primates12, it was not possible to solve, within the limit time, instances of the RDBMEP having a size comparable with the ones of the BMEP. For example, the exact algorithm for the RDBMEP was unable to tackle the instances M17, M18, RbcL55, M62, and M82 containing more than 12 taxa, while the exact algorithm for the BMEP (see [10]) tackled those instances up to 17, 18, 18, 19, and 17 taxa, respectively, in less than 1 h computing time. Similarly, the exact algorithm for the RDBMEP was unable to tackle the instances SeedPlant25, M43, and Rana64 containing more than 16, 16, and 15 taxa, respectively, versus 19, 29, and 20 taxa, respectively, relative to the exact algorithm for the BMEP. As for the solution time, the values of the gaps for the RDBMEP are also much higher than the corresponding ones relative to the BMEP. Specifically, gaps are usually higher then 30%, follow a trend that is inversely proportional to the uncertainty level, and approach 100% for an uncertainty level equal to 5%. This trend provides an insight of the harder nature of the RDBMEP with respect to the BMEP. In fact, in the BMEP the mainly difficulty is represented by the link between the evolutionary distances and the structural properties of the phylogenies (see [10]). These properties can be easily modeled, hence the linear relaxations of the formulations that embody them are usually characterized by very small gaps (less than 4% in average). Unfortunately, in the RDBMEP there exists a further difficulty represented by the dichotomous values that the distances can assume in their own intervals. Such aspect is much more difficult to characterize and relate to the structural properties of the phylogenies. Hence, the observed poor relaxations are possibly due to a lack of a systematic characterization of this aspect causes. Investigating such an issue is out of the scope of the present article and warrants additional analysis.

The number of the nodes expanded in the search tree has a trend proportional both to the increment of size of the instance and to the uncertainty level; however, it does not explode significantly as the presence of poor relaxations might suggest. This fact is possibly due to the SAS-like branching rules (discussed in Section 4) that, similarly to the BMEP, vastly help in breaking the high degree of symmetry of the problem. Finally, the results showed that the hardness of an instance increases by increasing the number of analyzed taxa and/or the uncertainty level. For example, the solution time increases by increasing the number of taxa, independently from the instance. Similarly, by doubling the uncertainty level from 5% to 10% prevents the exact algorithm from solving the instance M17 within the limit time. This phenomenon seems to be related to the number of equal entries in the distance matrix and to the level of overlap of the intervals. In fact, the higher the number of equal entries or overlapping intervals in the distance matrix the higher the number of equivalent optimal solutions to the problem. As result, the exact algorithm may be unable to prune fractional solutions in the search tree as they are characterized by equivalent relaxations. Developing strategies able to overcome such aspects could possibly speed up the solution times of the exact algorithm.

Table 1
Performances of the exact algorithm for the RDBMEP for different uncertainty levels in the input data.

Dataset	n	Uncertainty level 5%				Uncertainty level 10%				Uncertainty level 15%			
		Optimum	Time (s)	Gap (%)	Nodes	Optimum	Time (s)	Gap (%)	Nodes	Optimum	Time (s)	Gap (%)	Nodes
Primates12	10	0.001265	28.16	100	158	0.004248	46.09	100	238	0.00765	67.8	92.65	278
	11	0.001532	159.3	100	629	0.005277	344.06	100	1182	0.012475	1221.29	100	3021
	12	0.000563	182.7	100	423	0.004315	622.39	100	862	0.010935	5615.18	100	3052
M17	10	0.003044	1155.43	100	4 462	0.00643	2278.17	53	6290	0.011585	5185.82	32.17	8669
	11	0.003741	4726.81	100	11 485	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	12	0.003047	4976.01	100	8 525	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
M18	10	0.00277	172.26	100	859	0.009145	356.01	93.81	1225	0.016841	782.64	63.03	1857
	11	0.004029	924.41	100	2 833	0.011525	2245.96	100	3647	0.019855	6844.14	73.69	6038
	12	0.004408	3134.24	100	6 578	0.011973	4399.07	100	7642	n.a.	n.a.	n.a.	n.a.
SeedPlant25	10	0.001563	95.72	100	467	0.00472	118.91	100	661	0.007966	139.63	100	780
	11	0.001935	246.79	100	838	0.00555	289.54	100	1001	0.009006	340.03	100	1125
	12	0.001876	286.95	100	665	0.004976	333.15	100	768	0.008201	525.14	100	1124
	13	0.001988	1653.68	100	1 813	0.005359	2260.25	100	2474	0.00996	6678.9	100	4951
	14	0.001736	2047.73	100	2 010	0.005002	4246.2	100	4057	0.009419	10 389.48	100	7813
	15	0.001528	3882.11	100	2 640	0.004844	5837.71	100	3056	n.a.	n.a.	n.a.	n.a.
	16	0.002085	10 691.51	100	3 771	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
M43	10	0.001796	120.07	100	553	0.005722	246.39	58.43	809	0.010895	480.92	36.14	1028
	11	0.001779	579.86	100	1 781	0.005681	889.3	90.62	2437	0.010859	1937.05	52.41	3917
	12	0.001605	1208.89	100	2 580	0.005469	1903.23	95.76	2980	0.010903	6157.8	61.49	6299
	13	0.001615	761.45	100	1 043	0.005501	4633.7	100	5479	n.a.	n.a.	n.a.	n.a.
	14	0.001582	1346.17	100	1 365	0.005412	3969.14	100	3023	n.a.	n.a.	n.a.	n.a.
	15	0.002318	2636.17	100	2 037	0.001955	7404.73	100	3390	n.a.	n.a.	n.a.	n.a.
	16	0.001955	7404.73	100	3 390	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
RbcL55	10	0.005152	544.51	100	2 178	0.011977	1523.22	61.35	3622	0.020343	3813.93	42.73	5390
	11	0.00478	1203.54	100	3 443	0.010826	5544.49	78.78	5970	n.a.	n.a.	n.a.	n.a.
	12	0.004702	1870.59	100	3 805	0.011271	5567.2	100	6232	n.a.	n.a.	n.a.	n.a.
M62	10	0.003508	108.87	100	519	0.009268	190.03	55.99	720	0.017047	455.16	40.79	1437
	11	0.004193	310.17	100	1 025	0.011857	1625.07	73.04	3732	0.02078	4019.06	50.65	5651
	12	0.004703	879.52	100	1 916	0.014388	10 259.02	70.52	8617	n.a.	n.a.	n.a.	n.a.
Rana64	10	0.001088	97.28	100	562	0.002543	112.03	91.32	640	0.004205	174.44	63.28	883
	11	0.001043	448.64	100	1 591	0.00248	762.02	100	2666	0.004105	876.94	100	3087
	12	0.000955	570.15	100	1 352	0.002639	917.59	100	2193	0.004477	1693.12	100	3968
	13	0.001075	1378.47	100	2 093	0.002961	2148.07	100	3255	0.005215	4164.92	100	5988
	14	0.001349	3426.21	100	3 689	0.003265	5306.11	100	5928	0.005298	7545.86	100	7670
	15	0.001406	5571.7	100	4 383	0.003195	7729.42	100	6074	n.a.	n.a.	n.a.	n.a.
M82	10	0.001148	380.63	100	1 642	0.003019	676.66	100	2100	0.005375	1194.3	85.03	3340
	11	0.001493	1731.18	100	4 086	0.003374	2451.24	100	4199	0.005971	5928.97	86.45	6481
	12	0.001208	7886.1	100	19 107	0.002973	9470.83	100	21 076	n.a.	n.a.	n.a.	n.a.

As regards to the local search, we experienced a very good behavior of the heuristic approximate algorithm both in terms of the quality of the solutions provided and the solution times taken to solve the instances of the RDBMEP. Specifically, the local search took less than a few seconds to solve the considered instances of the RDBMEP and, in all cases in which it was possible to make a comparison, the local search returned the same optimal values returned by the exact algorithms. Due to these characteristics, we believe that the local search may constitute a valid alternative to perform phylogeny estimations of large molecular datasets affected by uncertainty.

7. Uncertainty: biological implications

Uncertainty is almost unavoidable in phylogeny estimation. It can occur in the dating process, in the form of noise in the measurements; in the sequencing process, in the form of sequencing errors; or in the computation of the evolutionary distances, in the form of under or over estimation of the dissimilarity of the involved taxa [5]. If the phylogenetic signal of taxa is strong enough, i.e., if related taxa tend to resemble each other with respect to their molecular sequences or phenotypic traits, then uncertainty may be negligible. This is the case e.g., for the mitochondrial DNA of primates (instance Primates12, [19]), where for uncertainty levels equal to 5%, 10%, and 15%, we did not observe any structural variation in the taxonomy of taxa (see Fig. 6). We need to push uncertainty to very high levels (e.g., 50%) before observing a variation, which in any case proves to be marginal (see e.g., Fig. 7: the presence of an uncertainty level of the 50% only changes the taxonomy of macaca fascicularis and macaca sylvanus).

In contrast, if the phylogenetic signal is not sufficiently strong, then the presence of uncertainty may have a major impact in recovering the phylogeny of taxa. This is the case e.g., for the mitochondrial DNA of Drosophilae (instance M17),

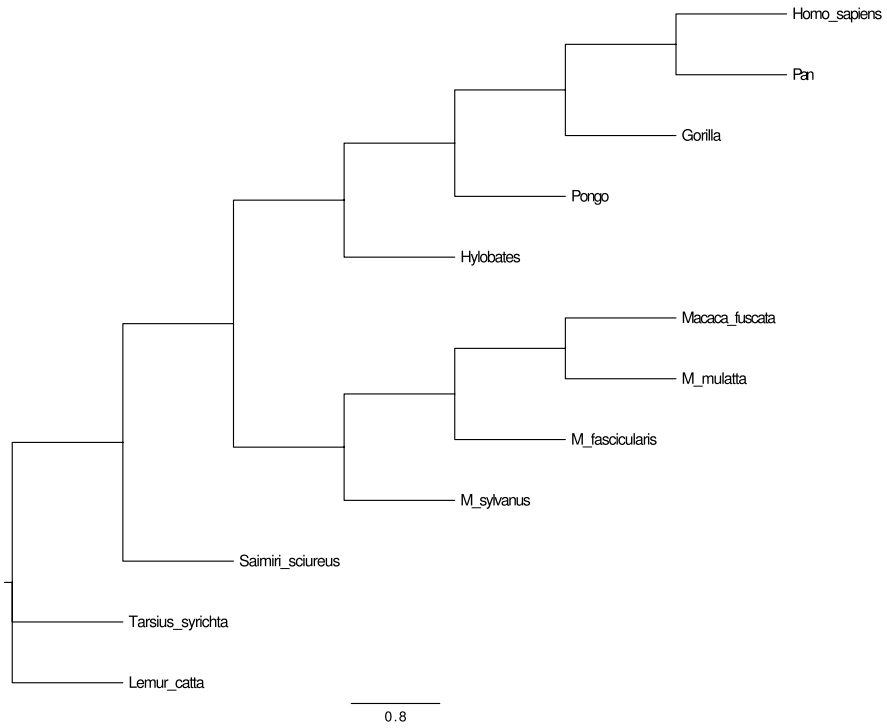

Fig. 6. The optimal balanced minimum evolution phylogeny to the instance Primates12.

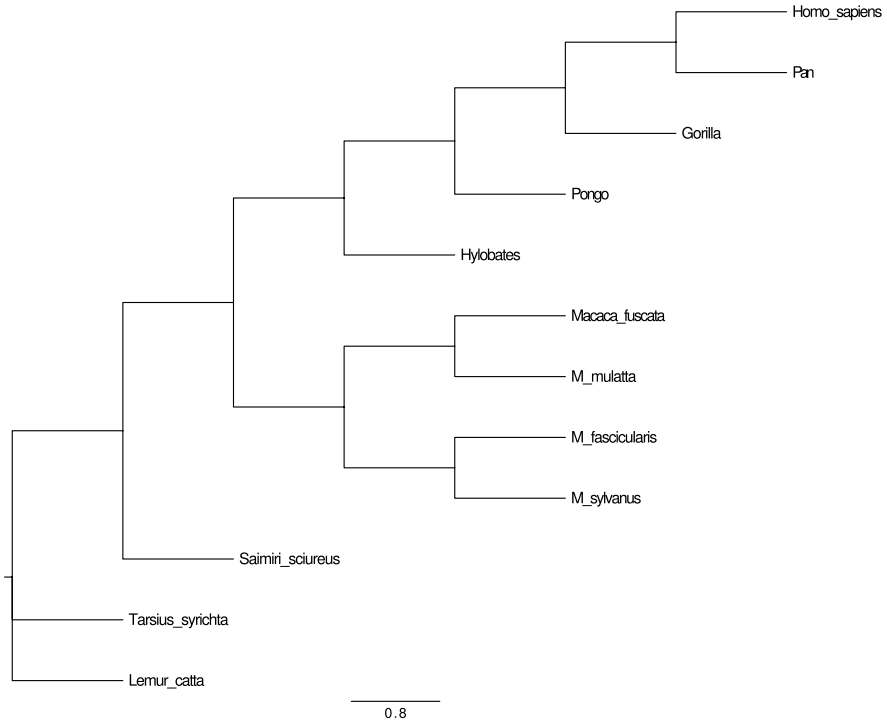

Fig. 7. The optimal robust deviation balanced minimum evolution phylogeny to the instance Primates12 with uncertainty level equal to 50%.

where we experienced important structural variations in the taxonomy of taxa already when considering uncertainty levels greater than or equal to 10% (see e.g., Figs. 8 and 9: the presence of an uncertainty level of 10% changes the taxonomy of the drosophila adunca, drosophila mimica, engiscaptomyza crassifemur, and drosophila melanogaster). In general, it is not

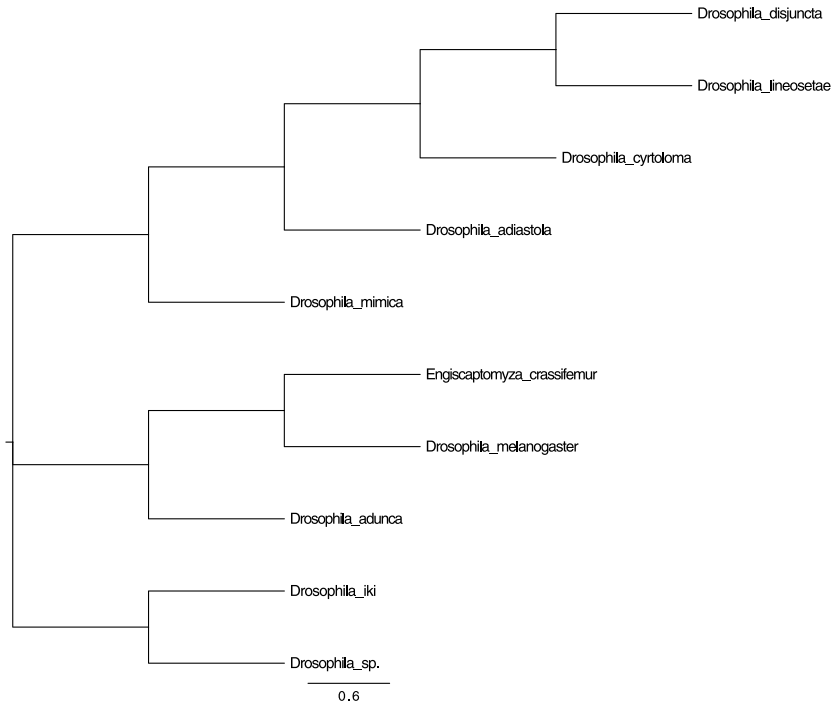

Fig. 8. The optimal balanced minimum evolution phylogeny to the instance M17 when considering the first 10 taxa.

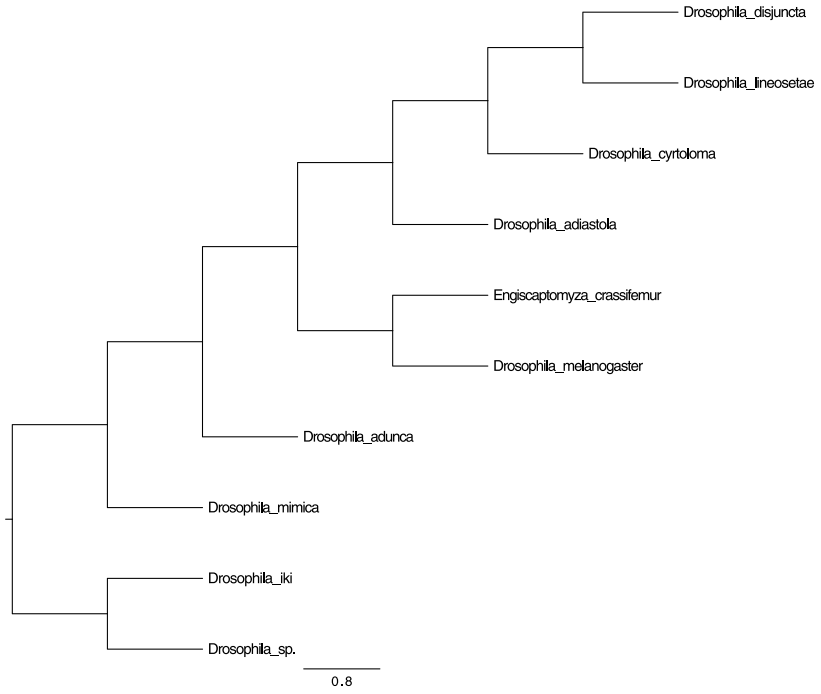

Fig. 9. The optimal robust deviation balanced minimum evolution phylogeny to the instance M15 when considering the first 10 taxa and an uncertainty level equal to 10%.

possible to know a priori whether the phylogenetic signal of the analyzed taxa is sufficiently strong or not, hence it is not possible to know a priori if the impact of uncertainty is negligible or not for the structure of the phylogeny. This is why having tools able to handle uncertainty in the input data is of fundamental assistance in phylogeny estimation.

8. Conclusion

The Balanced Minimum Evolution Problem (BMEP) is a recent version of the Phylogenetic Estimation Problem (PEP) [8] firstly introduced in [32]. Given a set Γ of n taxa and the corresponding matrix D of evolutionary distances, the BMEP consists of finding a phylogeny for Γ having minimum length [7]. The BMEP is based on the minimum evolution criterion of phylogenetic estimation, which states that if the evolutionary distances were unbiased estimates of the *true* evolutionary distances, i.e., the distances that one would obtain if all the molecular data from the analyzed taxa were available, then the true phylogeny would have an expected length shorter than any other possible phylogeny compatible with D. Interestingly, the minimum evolution criterion does not assert that molecular evolution follows minimum paths, but states, according to classical evolutionary theory, that a minimum length phylogeny may properly approximate the real phylogeny of well-conserved molecular data i.e., data whose basic biochemical functions undergone small change throughout evolution of the observed taxa [4]. Since the selective forces acting on taxa may not be constant over time, evolution proceeds by small rather than smallest change [4,36]. Thus, a minimum length phylogeny provides a lower bound on the overall number of mutation events that could have occurred along evolution of the observed taxa.

In this article we investigated for the first time the robust deviation balanced minimum evolution problem, i.e., a peculiar version the BMEP that arises whenever the evolutionary distances from taxa are uncertain and varying inside intervals. By exploiting some fundamental properties of its objective function, we presented a mixed integer programming model to exactly solve its instances and discussed the biological impact of uncertainty on the solutions to the problem. Our results give perspective on the mathematics of the RDBMEP and suggest new directions to tackle phylogeny estimation problems affected by uncertainty.

Acknowledgments

DC acknowledges support from the Belgian National Fund for Scientific Research (FNRS), of which he is "Chargé de Recherches". ML acknowledges support from the Communauté Française de Belgique—Actions de Recherche Concertées (ARC) and "Ministerio de Ciencia e Innovación" through the research project MTM2009-14039-C06. Finally, RP acknowledges support from the FNRS for the "bourse missions scientifiques". Part of this work was developed while RP was visiting the Graphs and Mathematical Optimization Unit of the Free University of Brussels. The authors also thank Dr. Patrick Mardulyn for helpful discussions and Dr. Rosa Maria Lo Presti for the datasets she provided.

References

[1] R. Aringhieri, D. Catanzaro, M. Di Summa, Optimal solutions for the balanced minimum evolution problem, Computers and Operations Research 38 (2011) 1845–1854.
[2] I.D. Aron, P. Van Hentenryck, On the complexity of the robust spanning tree problem with interval data, Operations Research Letters 32 (2004) 36–40.
[3] D.A. Bader, B.M.E. Moret, L. Vawter, Industrial applications of high-performance computing for phylogeny reconstruction, in: SPIE ITCom 4528, SPIE, Denver, CO, 2001, pp. 159–168.
[4] W.A. Beyer, M. Stein, T. Smith, S. Ulam, A molecular sequence metric and evolutionary trees, Mathematical Biosciences 19 (1974) 9–25.
[5] S.P. Blomberg, T. Garland, A.R. Ives, Testing for phylogenetic signal in comparative data: behavioral traits are more labile, Evolution 57 (4) (2003) 717–745.
[6] R.M. Bush, C.A. Bender, K. Subbarao, N.J. Cox, W.M. Fitch, Predicting the evolution of human influenza A, Science 286 (5446) (1999) 1921–1925.
[7] D. Catanzaro, The minimum evolution problem: overview and classification, Networks 53 (2) (2009) 112–125.
[8] D. Catanzaro, Estimating phylogenies from molecular data, in: R. Bruni (Ed.), Mathematical Approaches to Polymer Sequence Analysis and Related Problems, Springer, New York, 2011.
[9] D. Catanzaro, M. Labbé, R. Pesenti, J.J. Salazar-Gonzáles, Mathematical models to reconstruct phylogenetic trees under the minimum evolution criterion, Networks 53 (2) (2009) 126–140.
[10] D. Catanzaro, M. Labbé, R. Pesenti, J.J. Salazar-Gonzáles, The balanced minimum evolution problem, INFORMS Journal on Computing 24 (2) (2012) 276–294.
[11] D. Catanzaro, R. Pesenti, M. Milinkovitch, A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model, Bioinformatics 22 (6) (2006) 708–715.
[12] B.S.W. Chang, M.J. Donoghue, Recreating ancestral proteins, Trends in Ecology and Evolution 15 (3) (2000) 109–114.
[13] R. Desper, O. Gascuel, Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle, Journal of Computational Biology 9 (5) (2002) 687–705.
[14] R. Desper, O. Gascuel, Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle, Journal of Computational Biology 19 (5) (2002) 687–705.
[15] M. Farach, S. Kannan, T. Warnow, A robust model for finding optimal evolutionary trees, Algorithmica 13 (1995) 155–179.
[16] J. Felsenstein, Inferring Phylogenies, Sinauer Associates, Sunderland, MA, 2004.
[17] S. Fiorini, G. Joret, Approximating the balanced minimum evolution problem, Technical Report, Université Libre de Bruxelles, 2011.
[18] P.H. Harvey, A.J.L. Brown, J.M. Smith, S. Nee, New Uses for New Phylogenies, Oxford University Press, Oxford, UK, 1996.
[19] K. Hayasaka, T. Gojobori, S. Horai, Molecular phylogeny and evolution of primate mitochondrial DNA, Molecular Biology and Evolution 5 (1988) 626–644.
[20] O.E. Karasan, M.C. Pinar, H. Yaman, The robust shortest path problem with interval data, Technical Report, Bilkent University, Ankara, Turkey, 2001. http://www.optimization-online.org.
[21] P. Kouvelis, G. Yu, Robust Discrete Optimization and its Applications, Kluwer Academic Publishers, Boston, MA, 1997.
[22] V. Makarenkov, F.J. Lapointe, A weighted least-squares approach for inferring phylogenies from incomplete distance matrices, Bioinformatics 20 (13) (2004) 2113–2121.

[23] M.A. Marra, S.J. Jones, C.R. Astell, R.A. Holt, A. Brooks-Wilson, Y.S. Butterfield, J. Khattra, J.K. Asano, S.A. Barber, S.Y. Chan, A. Cloutier, S.M. Coughlin, D. Freeman, N. Girn, O.L. Griffith, S.R. Leach, M. Mayo, H. McDonald, S.B. Montgomery, P.K. Pandoh, A.S. Petrescu, A.G. Robertson, J.E. Schein, A. Siddiqui, D.E. Smailus, J.M. Stott, G.S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T.F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G.A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R.C. Brunham, M. Krajden, M. Petric, D.M. Skowronski, C. Upton, R.L. Roper, The genome sequence of the SARS-associated coronavirus, Science 300 (5624) (2003) 1399–1404.
[24] R. Montemanni, A benders decomposition approach for the robust spanning tree problem with interval data, European Journal of Operational Research 174 (2006) 1479–1490.
[25] R. Montemanni, J. Barta, M. Mastrolilli, L.M. Gambardella, The robust traveling salesman problem with interval data, Transportation Science 41 (3) (2007) 366–381.
[26] R. Montemanni, L.M. Gambardella, A branch and bound algorithm for the robust spanning tree problem with interval data, IDSIA-10-02, Technical Report, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, 2002.
[27] R. Montemanni, L.M. Gambardella, The robust shortest path problem with interval data via benders decomposition, 4OR 3 (2005) 315–328.
[28] C.Y. Ou, C.A. Ciesielski, G. Myers, C.I. Bandea, C.C. Luo, B.T.M. Korber, J.I. Mullins, G. Schochetman, R.L. Berkelman, A.N. Economou, J.J. Witte, L.J. Furman, G.A. Satten, K.A. Maclnnes, J.W. Curran, H.W. Jaffe, Molecular epidemiology of HIV transmission in a dental practice, Science 256 (5060) (1992) 1165–1171.
[29] L. Pachter, B. Sturmfels, The mathematics of phylogenomics, SIAM Review 49 (1) (2007) 3–31.
[30] F. Pardi, Algorithms on phylogenetic trees, Ph.D. Thesis, University of Cambridge, UK, 2009.
[31] D.S. Parker, P. Ram, The construction of Huffman codes is a submodular (convex) optimization problem over a lattice of binary trees, SIAM Journal on Computing 28 (5) (1996) 1875–1905.
[32] Y. Pauplin, Direct calculation of a tree length using a distance matrix, Journal of Molecular Evolution 51 (2000) 41–47.
[33] H.A. Ross, A.G. Rodrigo, Immune-mediated positive selection drives human immunodeficency virus type 1 molecular variation and predicts disease duration, Journal of Virology 76 (22) (2002) 11715–11720.
[34] M. Salazar-Neumann, Advances in robust combinatorial optimization and linear programming, Ph.D. Thesis, GOM—Université Libre de Bruxelles, 2010.
[35] M.S. Waterman, T.F. Smith, On the similarity of dendrograms, Journal of Theoretical Biololgy 73 (1978) 789–800.
[36] M.S. Waterman, T.F. Smith, M. Singh, W.A. Beyer, Additive evolutionary trees, Journal of Theoretical Biology 64 (1977) 199–213.
[37] H. Yaman, O.E. Karasan, M.C. Pinar, The robust spanning tree problem with interval data, Operations Research Letters 29 (2001) 31–40.